International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

The Theory from Large Deviations for Random Processes and Strong Convergence of Stochastic Approximation Procedures

Leonov, S.

IIASA Working Paper

WP-86-057

October 1986

The Theory of Large Deviations for Random Processes and Strong Convergence of Stochastic Approximation Procedures

S.L. Leonov

October 1986
WP-86-57

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

Preface

This paper deals with the application of "large deviation" theory to the analysis of stochastic approximation procedures. The approach allows to get new results in the asymptotical behaviour of stochastic procedures under very mild assumption about the "noise". The paper contains a short but illuminative survey of these results together with some new author's findings. For applications the last section seems to be interesting presenting some new ideas in multiobjective optimization.

Prof. V. Fedorov
Environmental Program

The Theory of Large Deviations for Random Processes and Strong Convergence of Stochastic Approximation Procedures

S.L. Leonov

1. INTRODUCTION

Stochastic recursive procedures or stochastic approximation procedures (SAP), as they are often called, take a significant role among the investigations on the probability theory and applied mathematics. Popularity of SAP may be explained by their relative simplicity (from computational point of view) and efficiency, as for a number of cases recursive procedures are asymptotically optimal.

Theoretical methods for the SAP's studying are in steady development. The "martingale" approach appeared to be the first fundamental approach for the analysis of recursive procedures. It is based on the application of the theory of martingales to construct a stochastic analogue of a Lyapunov function (for references see Nevel'son and Has'minskii, 1972).

During the last decade new results were obtained in the theory of random processes, concerning the analysis of the probabilities of large deviations (Ventsel and Freidlin, 1983). Application of the theory of large deviations (TLD) made it possible to find new results for the SAP's asymptotical behaviour. Recent monograph by Korostelev (1984) is devoted to the analysis of the necessary and sufficient conditions for almost sure (a.s.) convergence and rates of convergence in terms of upper functions. Kushner (1983, 1984) applies TLD to investigate asymptotics of the probabilities of escape time from the neighbourhood of stable points.

In the present paper an account of main ideas and results of the approach based on TLD is given. The range of problems is limited here by the studying of almost sure convergence for discrete-time procedures, main attention being given to the investigation of trajectories' behaviour in the neighbourhood of stable points.

The structure of the paper is the following. In section 2 we introduce basic notations and assumptions. Sections 3, 4 are devoted respectively to the investigation of convergence conditions and rates of convergence for SAP with independent noise. In section 5 SAP with dependent noise are analysed. In section 6 TLD is applied, as an illustrative example, to the problem of minimization of an additive function.

2. BASIC NOTATIONS AND ASSUMPTIONS

The following recursive procedure will be studied:

$$X(t+1) = X(t)+\alpha(t)[B(X(t)) + \varphi_t(X(t))] + \beta(t)\xi(t)\,, \qquad (1)$$

$t=1,2,\ldots$; $X(0) = x_0 \in R^d$ where x_0 is an arbitrary point in the Euclidean space R^d, $d \geq 1$; random vectors $\xi(t) = \xi(t,\omega)$ (the noise) are defined in the probability space (Ω,B,P) and have values in R^d . The set of assumptions is the follow-

ing.

A1. A point x_* is an asymptotically stable equilibrium point of the dynamic system

$$\dot{x}_s = B(x_s) \tag{2}$$

with the domain of attraction D , the condition

$$\sup_D \| B(x) \| < \infty$$

being held.

A2. Vector field $B(x): R^d \xrightarrow{B} R^d$ satisfies the Lipschitz condition in domain $D \setminus V_\varepsilon$ for every $\varepsilon > 0$, where V_ε is a ε-neighbourhood of x_* : for every $\varepsilon > 0$ there exists a positive constant L_ε , such that

$$\| B(x) - B(y) \| \le L_\varepsilon \| x - y \| \quad \text{for} \quad x, y \in D \setminus V_\varepsilon .$$

A3. With probability 1 (a.s.) trajectories $X(t) = X(t, \omega)$ return infinitely often into compact set K , $x_* \in K \subset D$.

A4. The sequences $\{\alpha(t)\}, \{\beta(t)\}$ are deterministic (non-random) sequences of positive scalars, and

$$\lim_{t \to \infty} \alpha(t) = \lim_{t \to \infty} \beta(t) = \lim_{t \to \infty} \frac{\beta^2(t)}{\alpha(t)} = 0, \sum_{t=1}^{\infty} \alpha(t) = \infty.$$

Function $\varphi_t(x)$ is a deterministic vector function: $R^d \xrightarrow{\varphi_t} R^d$, such that

$$\sup_{x \in D} \| \varphi_t(x) \| \xrightarrow[t \to \infty]{} 0$$

A5. $M \xi(t) = 0$ for all $t > 0$, $M \| \xi(t) \|^2 \le const < \infty$,

M is a symbol of averaging over measure P (we assume for simplicity that vectors $\xi(t)$ do not depend on space coordinate $X(t)$).

Now let us make some comments on introduced assumptions.

Assumption A2 makes it possible to consider functions $B(x)$ for which the Lipschitz condition is violated in x_* : e.g., in one-dimensional case ($d = 1$)

$$B(x) = -L \, | x - x_* |^P sign(x - x_*) + o(| x - x_* |^P), 0 \le P < 1, L > 0.$$

Assumption A3 is an analogue of the "boundedness condition" of Ljung (1977) and it helps to avoid traditional assumptions on behaviour of the field $B(x)$ as $\| x \| \to \infty$. It may be attained by introducing of a projector on K :

$$X(t+1) = \pi_K \{ Z(t) \},$$

where $Z(t)$ denotes right-hand side of (1).

It must be mentioned that several important optimization and estimation procedures may be put in the framework (1):

- the Robbins-Monro (RM) procedure (1951), $\alpha(t) = \beta(t)$, - an algorithm for finding the root of the equation $B(x) = 0$ when function $B(x)$ is measured with random error ξ (in the RM case we shall denote coefficients by $\gamma(t)$: $\alpha(t) = \beta(t) = \gamma(t)$),
- the Kiefer-Wolfowitz (KW) procedure (1952), $\frac{\alpha(t)}{\beta(t)} \xrightarrow[t \to \infty]{} 0$ - a stochastic analogue for a gradient method (in the KW case $\varphi_t(x)$ has a meaning of a deterministic bias of gradient's estimate),

- procedures of parametric adaptation (Tsypkin, 1971).

Throughout the paper the convergence of the procedure (1) means a.s. convergence of the trajectories $X(t)$ to a point x_* as $t \to \infty$.

3. CONVERGENCE OF SAP WITH INDEPENDENT NOISE

When we study SAP on the basis of TLD - approach the process $X(t)$, defined by (1), is considered as a small random perturbation of the stable dynamic system (2). Such an idea was proposed by Ljung (1974), Derevitskii and Fradkov (1974). The deviation of $X(t)$-trajectories from the trajectories of the dynamic system (2) may be estimated by means of the Gronwall-Bellman lemma:

Let $s(t)$ be a new time-variable (continuous time analogue),

$$s(t) = \sum_{1}^{t-1} \alpha(\tau),$$

Let $x(t)$ denote the solution of (2) at time $s(t)$ with initial value $x(N_0) = X(N_0)$. If $\sum_{N_0}^{N_1} \alpha(\tau) \le T$, $T > 0$, then there exists a positive constant $C = C(T)$, such that

$$\max_{N_0 \le t \le N_1} \|X(t) - x(t)\| \le C \max_{N_0 \le t \le N_1} \|\sum_{N_0}^{t} \beta(\tau)\xi(\tau)\| + o(1),\ o(1) \underset{N_0 \to \infty}{\to} 0 \qquad (3)$$

(independence of $\xi(t)$ is here of no importance).

Investigation of the conditions for a.s. convergence is based on the analysis of asymptotics for the probabilities of large deviations (PLD)

$$P\left\{\max_{N_i \le t \le N_{i+1}} \|Y_{N_i}(t)\| > \varepsilon\right\} \qquad (4)$$

for arbitrary fixed values $\varepsilon > 0$ and $T > 0$, where

$$Y_N(t) = \sum_{N}^{t} \beta(\tau)\xi(\tau),\ \sum_{N_i}^{N_{i+1}} \alpha(\tau) \sim T.$$

It follows from conditions

$$M\|\xi(t)\|^2 \le \text{const and } \lim_{t \to \infty} \frac{\beta^2(t)}{\alpha(t)} = 0$$

that the variance of random vectors $Y_{N_i}(t)$ tends to zero as $i \to \infty$, this fact explaining the origin of the term "large deviations".

By means of Levy's inequality estimation of PLD in (4) may be reduced to the estimation of the probability of deviation at the final time of interval $[N_i, N_{i+1}]: t = N_{i+1}$.

Levy's inequality (see Gihman and Skorohod, 1977, chapter 6).

Let $v_1, v_2, \cdots, v_n$ be independent random variables with zero mean, $\tilde{Y}(K) = \sum_{1}^{k} v_\tau$ and $\sqrt{M\tilde{Y}^2(n)} < \frac{\varepsilon}{2}$. Then

$$P\{\max_{1 \le k \le n} |\tilde{Y}(k)| < \varepsilon\} \le 2P\{|\tilde{Y}(n)| < \frac{\varepsilon}{2}\}. \qquad (5)$$

If we assume now that

$$\sum_t P\{\| Y_{N_t}(N_{t+1})\| > \varepsilon\} < \infty,$$

it is easy enough to yield the convergence of procedure (1). Indeed, it follows from (3) and the Borel-Cantelli lemma that for all sufficiently large t trajectories $X(t)$ a.s. will be close to the trajectories $x(t)$ of the dynamic system (2), and $x(t)$ converges to x, owing to stability assumptions.

It is well-known (see, for instance, Nevel'son and Has'minskii, 1972) that under condition

$$M\| \xi(t)| |^2 \le \text{const}$$

the sufficient condition for the convergence has the form

$$\sum_t \beta^2(t) < \infty. \tag{6}$$

If additional information on the distribution of $\{\xi(t)\}$ is known, then using exact estimates for the probabilities

$$P\{\| Y_{N_t}(N_{t+1})\| > \varepsilon\}, \tag{7}$$

it is possible to weaken condition (6) and even to obtain the necessary and sufficient convergence conditions for some cases.

Godovančuk and Korostelev (1983) considered the case of vectors $\xi(t)$ having power tails of distributions:

$$C_1 \le x^{\nu} P\{\| \xi(t)\| > x\} \le C_2\,,\ 0 < C_1 \le C_2 < \infty\,,\ \nu > 2 (x \ge \hat{x}) \tag{8}$$

and showed that the necessary and sufficient convergence condition has the form

$$\sum_t \beta^{\nu}(t) < \infty\,, \tag{9}$$

with application of the results by Hanson and Wright (1969) the probability in (7) is shown to have the order $o\left[\sum_{N_t}^{N_{t+1}} \beta^{\nu}(t)\right]$. The right-hand inequality in (8) is used to yield sufficiency, this inequality follows immediately from Chebyshev's inequality if the noise $\xi(t)$ has finite moment of order ν :

$$M\| \xi(t)| |^{\nu} \le \text{const}\ ;$$

necessity of (9) follows from the left-hand inequality in (8) and the Borel-Cantelli lemma.

Now let vectors $\xi(t)$ have finite exponential moment, i.e., Cramer's condition holds:

$$M \exp\{(Z, \xi(t))\} < \infty \text{ for all } t > 0\,,\ Z \in R^d \tag{10}$$

(the Gaussian noise or the bounded with probability 1 noise give appropriate examples). For this case Korostelev (1979) proved that the necessary and sufficient condition is the following:

$$\sum_t \alpha(t) \exp\{-\lambda \alpha(t) / \beta^2(t)\} < \infty \text{ for all } \lambda > 0. \tag{11}$$

For the RM procedure it has the form

$$\sum_t \exp\{-\lambda / \gamma(t)\} < \infty \text{ for all } \lambda > 0 \tag{11'}$$

(e.g., (11') holds with $\gamma(t) = t^{-c}, 0 < C \leq 1$) - here exponential estimates for PLD are used (Ventsel and Freidlin, 1970). The sufficiency of (11') for the Gaussian noise was earlier obtained by Ljung (1974).

4. RATES OF CONVERGENCE FOR SAP WITH INDEPENDENT NOISE

From the standpoint of the general theory of random processes convergent (in some probability sense) recursive procedures have a specific nature since their limiting distribution is degenerate. The analysis of rates of convergence is intended for finding a deterministic function $k(t)$, $k(t) \underset{t \to \infty}{\to} \infty$, such that the normalized process $Y(t)$,

$$Y(t) = k(t)[X(t) - x_*],$$

has a non-degenerate limiting distribution.

The first authors to study rates of convergence for SAP were Chung (1954), Sacks (1958) who analysed asymptotic normality of the normalized process and rates of moment's convergence. Methods based on the martingale approach received further development in the papers devoted to studying of a.s. rates of convergence for one-dimensional case ($d=1$). To formulate the results of these papers we will remind the definition of upper functions.

Definition. Let $X(t)$ a.s. converge to x_*. Deterministic sequence $h(t)$ is called an upper function for procedure (1) if

$$\overline{\lim_{t \to \infty}} \frac{X(t) - x_*}{h(t)} \overset{a.s.}{=} 1 .$$

Thus, $h(t)$ is a deterministic sequence which majorizes (for all sufficiently large t) random trajectories $X(t) - x_*$, this majorant being precise - it cannot be reduced.

Studying of upper functions is both of theoretical and practical interest: their application makes it possible to modify (1) so that $X(t)$ converges to x_* "from one side" (see, for instance, Anbar, 1977). Such modifications are useful for the problems with constraints on the domain of definition of function $B(x)$.

In papers by Gaposhkin and Krasulina (1974), Heyde (1974), Kersting (1977) it is shown that if function $B(x)$ is linear in the neighbourhood of x_* :

$$B(x) = -L(x - x_*) + o(x - x_*), \; L > 0,$$

and $\alpha(t) = \beta(t) = \frac{a}{t}$, $M\xi^2(t) = \sigma^2$, then upper function for procedure (1) is defined by the law of iterated logarithm:

$$h(t) = \sqrt{\frac{2a^2\sigma^2}{2aL - 1} \cdot \frac{\ln \ln t}{t}} \quad (2aL > 1). \tag{12}$$

The most strong result is obtained by Kersting - only the existence of variance is required for the noise $\xi(t)$. But linearity of function $B(x)$ is of primary importance for this method since it is based on the explicit solution of the linearized equation.

Application of the technique of TLD (Ventsel and Freidlin, 1983) makes it possible to obtain new results concerning rates of a.s. convergence for SAP. These results follow from the "sweeping theorem" (Korostelev, 1979): if condition (10) holds and the sum in (11) converges for $\lambda > \lambda_0$ and diverges for $\lambda < \lambda_0$ (e.g., $\alpha(t) = \beta(t) = \frac{\lambda_0}{t}$), then under some regularity of the coefficients $\alpha(t), \beta(t)$ the trajectories $X(t)$ a.s. have a set of limit points

$$W(\lambda_0)=\{x:V(x)\le\lambda_0\}$$

where

$$V(x)=\inf_{\varphi;T>0}\left\{\int_0^T H(\varphi(t)-B(\varphi(t)))dt\ ,\ \varphi(0)=x_*\ ,\ \varphi(T)=x\right\},$$

$\varphi(t)$ are absolutely continuous functions: $R^d \overset{\varphi}{\to} R^d$,

$H(u)$ is a Legendre transform of a function $G(Z) = \ln M \exp\{(Z,\xi(t)\}$:

$$H(u)=\sup_{Z\in R^d}[(u,Z)-G(Z)].$$

Applying the "sweeping theorem" Korostelev (1983) extended (12) to (a) multidimensional case and (b) non-linear functions $B(x)$ in one-dimensional case:

(a) if (10) holds, $\alpha(t)=t^{-1}$, $\beta(t)=t^{-b}$, $\frac{1}{2}<b\le 1$; matrix $COV=M[\xi(t)\xi^T(t)]$ is non-singular and $B(x)$ is linear in the neighbourhood of x_*:

$$B(x)=B_0(x-x_*)+\varphi(x-x_*),\ \|\varphi(x)\| / \|x\| \underset{\|x\|\to 0}{\to} 0,$$

where $B_0+I/2$ is a stable matrix (I - an identity ($d\times d$) - matrix), then the normalized process $Y(t)$,

$$Y(t)=(X(t)-x_*)\sqrt{t^{2b-1}/lnlnt}$$

a.s. has set G of limit points, where

$$\begin{cases} G=\left\{y \mid y^T S^{-1} y\le 2\right\} \\ S=\int_0^\infty e^{(B_0+I/2)s}\ COV\ e^{(B_0^T+I/2)s}\,ds \end{cases} \tag{13}$$

In particular, it follows from (13) that

$$\overline{\lim_{t\to\infty}}\frac{1}{2}Y^T(t)S^{-1}Y(t) \overset{a.s.}{=} 1$$

(it is well-known, see, for instance, Nevel'son and Has'minskii, 1972 - that finite-dimensional distributions of vectors $\sqrt{t^{2b-1}}(X(t)-x_*)$ converge, in the corresponding time-scale, to the distributions of the stationary Gaussian process with covariance matrix S . The above result shows that the set of limit points for the process $Y(t)$ coincides with the ellipsoid of equal probabilities for the limiting Gaussian distribution);

(b) in one-dimensional case, if (10) holds, upper functions are found for non-linear functions $B(x)$:

$$B(x)=-L\,|x-x_*|^P sign\ (x-x_*)+o(|x-x_*|^P)\ ,\ L>0\ ,\ P>0.$$

E.g., if $\alpha(t)=\beta(t)=t^{-C}$, $0<C\le 1$; $M\xi^2(t)=\sigma^2$, then upper functions are defined by the formula

$$h(t)=\left[\frac{\sigma^2(1+P-2Pc)}{2L}\ln\frac{t}{t^c}\right]^{1/(1+P)}\ ,\ \text{if}\ \frac{2PC}{1+P}<1 \tag{14}$$

(for $P\ne 1$, unlike the linear case, the solution of procedure (1) cannot be evaluat-

ed in the explicit form).

It must be noted that Cramer's condition of the existence of exponential moment (10) is rather restrictive, besides that, only the second moments of the noise $\xi(t)$ enter final results - namely covariance matrix COV in (13), variance σ^2 in (14). It turns out that condition (10) can be weakened.

Korostelev and Leonov (1983) studied the problem of existence of upper functions for SAP with the noise $\xi(t)$ having power tails of distributions (8). In multidimensional case (13) remains valid, and in one-dimensional case the form of upper functions coincides with (14) if condition

$$\frac{\nu PC}{1+P} > 1 \tag{15}$$

holds.

The fulfillment of condition (15) turns out to be of primary importance for the existence of upper functions. Let us consider, for example, a linear one-dimensional procedure

$$X(t+1) = X(t)+t^{-c}\cdot[-X(t)+\xi(t)]\,,\ 0 < C < 1, \tag{16}$$

where variables $\xi(t)$ have power tails of distributions (8). Let condition (15), which here has the form $\nu C > 2$, be violated and instead the following condition holds:

$$1 < \nu C < 2 \tag{17}$$

(the left-hand inequality provides that $X(t)$ a.s. converges to a point $x_* = 0$ - cf. with (9)).

The process $t^{c/2}X(t)$ is asymptotically normal, but under condition (17) there does not exist an upper function for procedure (16) in the class of non-increasing sequences $h(t)$ (see Korostelev and Leonov, 1983):

$$\text{if } \sum_t \left[h(t)t^c\right]^{-\nu} < \infty\,, \text{ then } \lim_{t\to\infty} h^{-1}(t)X(t) \overset{a.s.}{=} 0.$$

$$\text{if } \sum_t \left[h(t).t^c\right]^{-\nu} = \infty\,, \text{ then } \overline{\lim_{t\to\infty}}\, |h^{-1}(t)X(t)| \overset{a.s.}{=} +\infty.$$

This example is apparently of origin interest for the theory of random processes.

5. SAP WITH DEPENDENT NOISE

While investigating SAP with independent noise the main attention, as shown above, is given to the analysis of interdependence between the conditions for the coefficients $\alpha(t)$, $\beta(t)$ and for the distribution of random vectors $\xi(t)$. In the case of dependent noise it is necessary to introduce the conditions for weakening of dependence between elements of the random sequence $\{\xi(t)\}$ (so that the "past" and the "future" are asymptotically independent). Traditionally the dependence of the $\{\xi(t)\}$-sequence is described by different mixing conditions (see Rosenblatt, 1974, or Ibragimov and Linnik, 1965).

The first publications on SAP with dependent noise appeared in the beginning of the 1960s (Driml and Nedoma, 1960; Sakrison, 1964). During the last decade the number of papers devoted to this problem significantly increased, the methods for the analysis of SAP with independent noise being developed and extended to the dependent case. The approaches for studying of SAP with dependent noise are diverse. They are based on the theory of quasimartingales (Borodin, 1979), the averaging principle (Geman, 1979; Krasulina, 1975; Kul'chitskii, 1978), the

analysis of conditional means (Poznyak and Chikin, 1984); also such publications may be mentioned as Ljung (1977, 1978), Kushner (1977, 1978), Farden (1981), Solo (1982). Bibliography may be found in Korostelev and Leonov (1984), Poznyak and Chikin (1984).

Kushner (1983, 1984) applies TLD to investigate asymptotics of the probabilities of escape time from the neighbourhood of stable points of SAP with dependent noise.

Application of the TLD-approach for the convergence analysis of SAP with dependent noise makes it possible to obtain the results which are similar to the corresponding results in the independent case or even coincide with them. In this section we will confine ourselves to the formulation of the results for the RM-type procedures ($\alpha(t)=\beta(t)=\gamma(t)$).

Leonov (1982) (see also Korostelev, 1984, chapter 6) studied the necessary and sufficient convergence conditions for procedure (1) where vectors $\xi(t)$ have power tails of distribution (8) and satisfy strong mixing condition:

$$\rho(t) = \sup_{F_1 \in F_{\le s}, F_2 \in F_{\ge s+t}} |P(F_2) - P(F_2|F_1)| \underset{t\to\infty}{\longrightarrow} 0,$$

$F_{\ge s}, F_{\le s}$ are sigma-algebras, generated by sequences $\{\xi(s),\xi(s+1),\ldots\}$ and $\{\xi(1),\cdots,\xi(s)\}$ respectively, $\rho(t)$ is a coefficient of strong mixing (see Ibragimov and Linnik, 1965). If condition

$$\sum_t \rho(t) < \infty \tag{18}$$

holds, then it is sufficient for the convergence that there exists an arbitrary small positive δ , such that

$$\sum_t \gamma^{\nu-\delta}(t) < \infty. \tag{19}$$

If coefficient $\rho(t)$ exponentially decreases:

$$\rho(t) \le Ce^{-\lambda t} \quad \text{for some} \quad C > 0, \lambda > 0, \tag{20}$$

then the necessary condition is the following:

$$\sum_t \gamma^{\nu+\delta}(t) < \infty \quad \text{for all} \quad \delta > 0 \tag{21}$$

(similar to the independent case, the right-hand inequality in (8) is used to obtain sufficiency, the left-hand one - to obtain necessity).

While proving the sufficiency of condition (19) the key role is played by the analogues of Levy's inequality (5) and Hanson-Wright's estimate for dependent noise. Here is the formulation of these results.

Lemma. Let condition (18) hold. Then for arbitrary $\varepsilon>0$, $T>0$, $\delta>0$ there exist constants $C_1=C_1(\varepsilon,T)>0, C_2=C_2(\varepsilon,T,\delta)>0$ and natural number $\tau=\tau(\varepsilon,T)$, such that

$$P\left\{\max_{N_0\le t\le N_1} \|Y_{N_0}(t)\| > \varepsilon\right\} \le C_1\left[P\left\{\|Y_{N_0}(N_1)\| \ge \frac{\varepsilon}{4}\right\} + \sum_{N_0}^{N_1+\tau} \gamma^{\nu}(t)\right., \tag{22}$$

$$P\{\|Y_{N_0}(N_1)\| > \varepsilon\} \le C_2 \sum_{N_0}^{N_1} \gamma^{\nu-\delta}(t), \tag{23}$$

inequalities (22), (23) being valid uniformly for such N_0, N_1 , that

$$\sum_{N_0}^{N_1} \gamma(t) \le T \ (Y_{N_0}(t) = \sum_{N_0}^{t} \gamma(\tau)\xi(\tau)).$$

Here is an ***example***: Let $\eta(t)$ be i.i.d. random variables with density of distribution $p(x)=C_0(1+|x|^{1+\nu})^{-1}$, C_0-normalizing constant, then sequence $\{\xi(t)\}$, defined by

$$\xi(t+1)=f(\xi(t))+\eta(t)\ ,\ \xi(o)=0,$$

$$f(\xi)=\left\{\begin{matrix}\xi, |\xi|\le A\\ A\,\mathrm{sign}\xi, |\xi|>A\end{matrix}\right\},\ A>0$$

($f(\xi)$ is a saturation function), satisfies (8), (20) and thus, also (18) - it can be easily shown that sequence $\xi(t)$ satisfies Doeblin's condition (see Doob, 1953).

Another example of the noise satisfying strong mixing condition is a sequence of m-dependent random vectors: $p(t)\equiv 0$, $t>m$. By means of inequality (22) the necessary and sufficient convergence condition is obtained for this case and it has the form

$$\sum_t \gamma^{\nu}(t)<\infty, \tag{24}$$

which coincides with independent case (cf. with (9)).

Strong mixing condition is rather restrictive and eliminates from consideration some interesting examples. Such an example is provided by a sequence $\{\xi(t)\}$ having "innovations representation":

$$\left\{\begin{matrix}\xi(t)=\sum\limits_{k=-\infty}^{t} h(t,k)\eta(k)\\ |h(t,k)|\le const.\,\lambda^{t-K}, 0<\lambda<1;\ |h(t,t)|\ge h_0>0\end{matrix}\right., \tag{25}$$

where $\eta(k)$ are independent random vectors with zero mean. SAP with the noise having the representation (25) were studied by Ljung (1978). Exact estimates for PLD are not used by Ljung, but rough estimates like Chebyshev's inequality, and in our notation his result may be formulated in the following manner: if for some even number $\nu, \nu\ge 4$,

$$M(\|\eta(t)\|^{\nu})\le const\ <\infty,$$

$\gamma(t)$ is a non-increasing sequence, such that

$$\overline{\lim_{t\to\infty}}\left[\frac{1}{\gamma(t+1)}-\frac{1}{\gamma(t)}\right]<\infty, \tag{26}$$

then condition

$$\sum_t \gamma^{\nu/2}(t)<\infty \tag{27}$$

is sufficient for the convergence (condition (27) is, certainly, less restrictive than the traditional condition (6)).

With the help of the analysis of exact asymptotics for PLD Korostelev and Leonov (1984) strengthened Ljung's result: if in (25) vectors $\eta(k)$ have power tails of distributions (8), then the necessary and sufficient convergence condition has the form (24) and is the same as for SAP with independent noise.

An important example of the noise satisfying (25) is a stationary autoregressive scheme ($\alpha=1$):

$$\sum_{i=0}^{q}\xi(t-i)a_i=\eta(t), \tag{28}$$

where q is an arbitrary natural number, a_i are real constants, such that all the

roots $x_{(i)}$ of the characteristic equation

$$\sum_{i=0}^{q} x^i a_{q-i} = 0$$

satisfy inequality: $|x_{(i)}|<1$, $i=\overline{1,n}$.

The case of dependent noise, forming Markov chain on the compact set and thus, satisfying condition (10), was studied by Korostelev (1984, chapter 6). It is shown that under regularity conditions of type (26) the necessary and sufficient convergence condition coincides with the one for the independent case and has the form (11). The "sweeping theorem" is generalized also for the case of dependent noise having finite exponential moment (10).

With the application of limit theorems by Ventsel (1979) and Freidlin (1978) some results are obtained concerning rates of convergence for SAP with dependent noise having power tails of distributions and forming a stationary autoregressive scheme (28). If condition (15) holds, random variables $\eta(t)$ satisfy (8) and $M\eta^2(t)=\sigma_0^2$, then upper functions are defined by formula (14), where

$$\sigma^2=\sigma_0^2/(\sum_{0}^{q} a_i)^2$$

(see Leonov, 1984). The result, absolutely analogous to the independent case, is also true for the procedure (16) with autoregressive noise.

6. A STOCHASTIC RECURSIVE PROCEDURE FOR MINIMIZATION OF AN ADDITIVE FUNCTION

In this section the theory of large deviations is applied, as an illustrative example, to the analysis of a recursive procedure for minimization of an additive function:

$$F_p(x)=\sum_{i=1}^{n} p_i f_i(x) \to \min_{x\in R^d} ; p_i\geq 0 , \sum_i p_i=1.$$

Procedure under consideration has the following form:

$$X(t+1)=\pi_{U_R}\{X(t)-\gamma(t)\nabla f_{\vartheta(t)}(X(t))\},X(0)\in R^d, \tag{29}$$

where $U_R=\{x:\|x\|\leq R\}$, π_{U_R} is a projector.

Algorithms of type (29) may be used when the application of traditional optimization methods, such as gradient methods, meets with certain difficulties - e.g., n is large enough or calculation of functions $f_i(x)$ is laborious. In such situations it is possible to use functions $\varphi(x)$ one by one for finding the minimum

$$x_*^{(p)}=\operatorname*{argmin}_{R^d} F_p(x),$$

indexes $\vartheta(t)$ being selected in a deterministic ($\vartheta(t)=t\,[mod n]$, for example) or a random manner.

Here we shall study procedure (29) with random selection of indexes: $\{\vartheta(t)\}$ is a sequence of i.i.d. random variables, such that

$$P\{\vartheta(t)=i\}=p_i , i=\overline{1,n}.$$

The following assumptions are also supposed to be fulfilled.

B1. $\{\gamma(t)\}$ is a deterministic sequence of positive numbers, it satisfies traditional conditions

$$\lim_{t\to\infty}\gamma(t)=0 , \sum_t \gamma(t)=\infty,$$

and a "regularity condition": for every $T>0$

$$\lim_{t_0\to\infty}\left[\max_{t_0\le t\le t_1}\gamma(t)\right]/\left[\min_{t_0\le t\le t_1}\gamma(t)\right]=1\ ,\ \text{if}\ \sum_{t_0}^{t_1}\gamma(t)\le T \tag{30}$$

(e.g., (30) is satisfied with $\gamma(t)=t^{-\alpha}$, $0<\alpha<1$).

B2. $f_i(x), i=\overline{1,n}$, are convex continuous differentiable functions: $R^d \overset{f_i}{\to} R^1$; $F_p(x)$ has a unique minimum for arbitrary set $\{p_i\}$, such that

$$p_i \ge 0, \sum_i p_i = 1.$$

Moreover, there exist integers j and m , such that

$$x_j^* \ne x_m^* \ , \ x_j^* = \operatorname*{argmin}_{R^d} f_j(x).$$

B3. Gradients $\{\nabla f_i(x)\}$ satisfy the Lipschitz condition in U_R and $x_i^* \in U_R$, $i=\overline{1,n}$; R is a positive constant, such that

$$(\nabla f_i(x), x) > 0, \|x\| > R (i=1,n)$$

(existence of such R follows from B2).

Procedure (29) can be put in the framework (1):

$$X(t+1)=\pi_{U_R}\{X(t)+\gamma(t)[B(X(t))+\xi_{X(t)}(t)]\},$$

where $B(x)=-\sum_i p_i \nabla f_i(x); P\{\xi_x(t)=-\nabla f_i(x)-B(x)\}=p_i$

(here the noise $\xi_x(t)$ depends on space coordinate x). Thus, the main results of the SAP's analysis discussed in sections 3, 4 remain valid. Since vectors $\xi_x(t)$ are bounded with probability 1, i.e., condition (10) holds, the necessary and sufficient condition for the convergence to a point $x_*^{(p)}$ has the form (11). The analogue of (13) also remains true, matrix COV having representation

$$COV = M\left[\xi_{x_*^{(p)}}(t).\xi^T_{x_*^{(p)}}(t)\right] = \sum_{i=1}^{n} p_i \left[\nabla f_i(x_*^{(p)})\right]\left[\nabla f_i(x_*^{(p)})\right]^T.$$

If the sum in (11') converges for $\lambda>\lambda_0$ and diverges for $\lambda<\lambda_0$, then there exists a set $W_p(\lambda_0)$ of a.s. limit points for the trajectories $X(t)$:

$$W_p(\lambda_0) = \{x \in R^d : V_p(x) \le \lambda_0\},$$

where

$$V_p(x) = \inf_{T>0}\{I_{0,T}(x_*^{(p)}, \varphi), \varphi(T) = x\},$$

$$I_{0,T}(y, phi) = \int_0^T H(\varphi(t), \dot\varphi(t) - B(\varphi(t))dt, \varphi(0) = y,$$

$\varphi(t)$ are absolutely continuous functions,

$H(x,u)$ is a Legendre transform of a function $G_x(Z)$:

$$G_x(Z) = \ln M \exp\left[Z, \xi_x(t))\right], \ H(x,u) = \sup_Z \left[u, Z) - G_x(Z)\right].$$

Now let us consider procedure (29) for which the sum in (11') diverges for all $\lambda>0$: $\lambda_0=\infty$, e.g., $\gamma(t)=1/\sqrt{lnt}$. In this case the set of limit points has the representation

$W_p = W_p(\infty) = \{x$: there exist an absolutely continuous function $\varphi(t)$ and positive T, such that $\varphi(0) = x_*^{(p)}$, $\varphi(T) = x$ and $I_{0,T}(x_*^{(p)}, \varphi) < \infty\}$.

The following theorem establishes the relation between W_p and Q_f, where Q_f is a set of Pareto-optimal solutions of a multicriterial minimization problem

$$\{f_1(x), \ldots, f_n(x)\} \to \min_{x \in R^d}$$

(it must be outlined that minimization of the weighted sum $\sum_i p_i f_i(x)$ is one of the ways for criteria's scalarization).

Let $\bar{W}_p$ be a closure of W_p.

Theorem. $\bar{W}_p$ does not depend on $\{p_i\}$: $\bar{W}_p \equiv W$, where W is a bounded closed set, such that

$$W \supseteq Q_f.$$

Moreover, Let E be a domain of accessibility for the dynamic system

$$\dot{\varphi}(t) = -\sum_i v_i(t) \nabla f_i(\varphi(t)), \varphi(0) = x_i^* \tag{31}$$

$v_i(t)$ are absolutely continuous non-negative functions,

$$\sum_i v_i(t) \equiv 1.$$

Then $\bar{E} = W$, where $\bar{E}$ is a closure of E

($E = \{x$: there exist a function $\varphi(t)$ and t_1, such that $\varphi(t)$ satisfies (31) and $\varphi(t_1) = x$).

The proof of the theorem is based on the TLD-technique (Gartner, 1977; Ventsel and Freidlin, 1983) and is obtained by the present author (Leonov, 1985). This publication also contains examples of sets W, Q_f in the Euclidean space R^2. For some cases these sets coincide: $W = Q_f$, for another strict imbedding takes place: $W \supset Q_f$.

REFERENCES

Anbar, D. (1977): A modified Robbins-Monro procedure approximating zero of a regression function from below. - Ann. Statist., v.5, No. 1, pp. 229-234.

Borodin, A.N. (1979): A stochastic approximation procedure in the case of weakly dependent observations. - Theor. Prob. Appl., v. 24, No. 1.

Chung, K.L. (1954): On a stochastic approximation method. - Ann. Math. Stat., v. 25, No. 3, pp. 463-483.

Derevitskii, D.P. and A.L. Fradkov (1974): Two models for analyzing the dynamics of adaptation algorithms. - Automat. Telemech., No. 1 (Automat. and Remote Contr., v. 35).

Driml, M. and J. Nedoma (1960): Stochastic approximation for continuous random processes. - In: Trans. 2nd Prague Conference on Inform. Theory, Stat. Decision Functions and Random Proc., Prague, pp. 145-158.

Doob, J.L. (1953): Stochastic Processes. - John Wiley & Sons, N.Y.

Farden, D.C. (1981): Stochastic approximation with correlated data. - IEEE Trans. on Information Theory, v. IT-21, No. 1, pp. 105-113.

Freidlin, M.I. (1978): The averaging principle and theorems on large deviations. - Russian Math. Surveys, v. 33, July-Dec.

Gaposhkin, V.F. and T.P. Krasulina (1974): On the law of iterated logarithm for stochastic approximation procedures. - Theor. Prob. Appl., v. 19, No. 4.

Gartner, J. (1977): On large deviations from the invariant measure. - Theor. Prob. Appl., v. 22, No. 1.

Geman, S. (1979): A method of averaging for random differential equations with applications to stability and stochastic approximation. - In: Approximate Solution of Random Evolution Equations, N.Y. - Oxford: North Holland, pp. 49-86.

Gihman, I.I. and A.V. Skorohod (1977): Introduction to the Theory of Random Processes. - Moscow: Izd. Nauka (in Russian).

Godovančuk, V.V. and A.P. Korostelev (1983): Conditions for the local convergence of recursive stochastic procedures. - Theor. Prob. Appl., v. 28, No. 1.

Hanson, D.L. and F.T. Wright (1969): Some more results on rates of convergence in the law of large numbers for weighted sums of independent random variables. - Trans. Amer. Math. Soc., v. 141, pp. 443-464.

Heyde, C.C. (1974): On Martingale limit theory and strong convergence results for stochastic approximation procedures. - Stoch. Proc. Appl., v. 2, No. 4, pp. 359-370.

Ibragimov, I.A. and Yu. V. Linnik (1965): Independent and Stationary Random Variables. - Moscow: Izd. Nauka (in Russian).

Kersting, G. (1977): Almost sure approximation of the Robbins-Monro process by sums of independent random variables. - Ann. Prob., v. 5, No. 6, pp. 954-965.

Kiefer, J. and J. Wolfowitz (1952): Stochastic estimation of the maximum of a regression function. - Ann. Math. Stat., v. 23, No. 3, pp. 462-466.

Korostelev, A.P. (1979): Damping perturbations of dynamic systems and convergence conditions for recursive stochastic procedures.
- Theor. Prob. Appl., v. 24, No. 2.

Korostelev, A.P. (1983): A note on upper functions for stochastic approximation. - Theor. Prob. Appl., v. 28, No. 4.

Korostelev, A.P. (1984): Stochastic Approximation Procedures (Local Properties). - Moscow: Izd. Nauka (in Russian).

Korostelev, A.P. and S.L. Leonov (1983): Upper functions for stochastic approximation procedures with power tails of distributions. - In: Dynamics of Non-Homogeneous Systems, Moscow: VNIISI, pp. 55-64 (in Russian).

Korostelev, A.P. and S.L. Leonov (1984): Stochastic approximation procedures with stationary noise. - In: Statistical Models and Methods, Moscow: VNIISI, pp. 46-61 (in Russian).

Krasulina, T.P. (1975): Some notes on the stochastic approximation. - Automat. Telemech., No. 7.

Kul'chitskii, O.Yu. (1978): Non-Markov algorithms for statistical optimization with continuous time. - Automat. Telemech., No. 5,6.

Kushner, H.J. (1977): General convergence results for stochastic approximations via week convergence theory. - J. Math. Anal. Appl., v. 61, pp. 490-503.

Kushner, H.J. and D.S. Clark (1978): Stochastic Approximation for Constrained and Unconstrained Systems. - Springer.

Kushner, H.J. (1983): Asymptotic behaviour of stochastic approximation and large deviations. - LCDS Report 83-1, Brown Univ., Providence, R.I.

Kushner, H.J. and P. Dupuis (1984): Stochastic approximations via large deviations: asymptotic properties. - LCDS Report 84-2, Brown Univ., Providence, R.I.

Leonov, S.L. (1982): Conditions for the convergence of stochastic approximation procedures with dependent noise. - In: Data Analysis in System Studies, Moscow: VNIISI, pp. 44-54 (in Russian).

Leonov, S.L. (1984): Asymptotic properties of stochastic recursive procedures with autoregressive noise. - In: Methods of Complex Systems Studying, Moscow: VNIISI, pp. 20-27 (in Russian).

Leonov, S.L. (1985): A stochastic algorithm for minimization of an additive function. - Automat. Telemech., No. 10.

Ljung, L. (1974): Convergence of recursive stochastic algorithms. - Report 7403, Division of Automatic Control, Lund Institute of Technology, Lund, Sweden.

Ljung, L. (1977): Analysis of recursive stochastic algorithms. - IEEE Transactions on Automatic Control, v. AC-22, No. 4, pp. 551-575.

Ljung, L. (1978): Strong convergence of a stochastic approximation algorithm. - Ann. Stat., v. 6, No. 3, pp. 680-696.

Nevel'son, M.B. and R.Z. Has'minskii (1972): Stochastic Approximation and Recursive Estimation. - Moscow: Izd. Nauka (in Russian).

Poznyak, A.S. and D.O. Chikin (1984): Asymptotical properties of stochastic approximation procedures with dependent noise. - Automat. Telemech., No. 12.

Robbins, H. and S. Monro (1951): A stochastic approximation Method. - Ann. Math. Stat., v. 22, No. 2, pp. 400-407.

Rosenblatt, M. (1974): Random Processes. - Springer (2nd ed.)

Sacks, J. (1958): Asymptotic distribution of stochastic approximation procedures. - Ann. Math. Stat., v. 29, No. 2, pp. 373-405.

Sakrison, D.J. (1964): A continuous Kiefer-Wolfowitz procedure for random processes. - Ann. Math. Stat., v. 35, No. 2, pp. 590-599.

Solo, V. (1982): Stochastic approximation with dependent noise. - Stoch. Proc. Appl., v. 13, pp. 157-170.

Tsypkin, Ya. Z. (1971): Adaptation and Learning in Automatic Systems. - N.Y.: Academic.

Ventsel, A.D. (1976): Rough limit theorems on large deviations for Markov processes, I, II. - Theor. Prob. Appl., v. 21, No. 2, 3.

Ventsel, A.D. (1979): Rough limit theorems on large deviations for Markov processes, III. - Theor. Prob. Appl., v. 24, No. 4.

Ventsel, A.D. and M.I. Freidlin (1970): On small random perturbations of dynamic systems. - Russian Math. Surveys, v. 25, Jan.-June.

Ventsel, A.D. and M.I. Freidlin (1983): Large Deviations. - Springer.